# SHORT COMMUNICATION

# The Histone Database: A Comprehensive Resource for Histones and Histone Fold-Containing Proteins

**Leonardo Mariño-Ramírez,[1] Benjamin Hsu,[2] Andreas D. Baxevanis,[2] and David Landsman[1]***

[1]*Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland*

[2]*Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland*

***ABSTRACT*** **The Histone Database is a curated and searchable collection of full-length sequences and structures of histones and nonhistone proteins containing histone-like folds, compiled from major public databases. Several new histone fold-containing proteins have been identified, including the huntingtin-interacting protein HYPM. Additionally, based on the recent crystal structure of the Son of Sevenless protein, an interpretation of the sequence analysis of the histone fold domain is presented. The database contains an updated collection of multiple sequence alignments for the four core histones (H2A, H2B, H3, and H4) and the linker histones (H1/H5) from a total of 975 organisms. The database also contains information on the human histone gene complement and provides links to three-dimensional structures of histone and histone fold-containing proteins. The Histone Database is a comprehensive bioinformatics resource for the study of structure and function of histones and histone fold-containing proteins. The database is available at http://research.nhgri.nih.gov/histones/. Proteins 2006;62:838–842.** © 2005 Wiley-Liss, Inc.*

**Key words: histones; histone-like proteins; multiple sequence alignments; Histone Database**

## INTRODUCTION

Histone proteins have central roles in both chromatin organization (as structural units of the nucleosome) and gene regulation (as dynamic components that have a direct impact on DNA transcription and replication).[1] Eukaryotic DNA wraps around a histone octamer to form a nucleosome, the first order of compaction of eukaryotic chromatin.[1] The core histone octamer is composed of a central H3-H4 tetramer and two flanking H2A-H2B dimers.[2] Each of the four core histones contains a common structural motif, called the histone fold, which facilitates the interactions between the individual core histones. The histone fold is composed of three α-helices connected by two loops, which allow heterodimeric interactions between core histones known as the "handshake" motif.[3] Although each individual histone protein family is highly conserved, the histone fold is not conserved at the level of sequence; despite this, the structures of these proteins are conserved.[4] A higher-resolution crystal structure of the nucleosome core particle demonstrated a more detailed structure of the histone folds in each of the histones.[5] In addition to the core histones, there is a "linker histone" called H1 (or H5 in avian species). The linker histones, which do not contain the histone fold motif, are critical to the higher-order compaction of chromatin, because they bind to internucleosomal DNA and facilitate interactions between individual nucleosomes (reviewed in Bustin et al.[6]). In addition, H1 variants have been shown to be involved in the regulation of developmental genes.[7]

Histone proteins and their variants have critical roles in gene regulation. Recently, it has been shown that nucleosomes are disassembled at transcriptionally active promoters.[8,9] Core histones can also have a variety of posttranslational modifications that have a role in the transition between transcriptionally active or silent chromatin.[10] The core histones, having their origins in *Archaea*, are among the most slowly evolving eukaryotic proteins. However, over evolutionary time, members of the histone H2A and H3 families have assumed specialized roles in DNA repair, gene silencing, gene expression, and centromere function.[11] The histone fold motif has also been found in a variety of nonhistone proteins, including the NF-Y transcription factor and the Ras activator Son of Sevenless.[4] Recently, the structure of the N-terminal region of Son of Sevenless was solved and the presence of two tandem histone folds was confirmed at the structural level.[12]



Grant sponsor: Intramural Research Program of the National Institutes of Health, National Library of Medicine.

*Correspondence to: David Landsman, Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, MSC 6075, Bethesda, MD 20894-6075.
E-mail: landsman@ncbi.nlm.nih.gov

Received 24 May 2005; Revised 12 September 2005; Accepted 20 September 2005

Published online 12 December 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20814

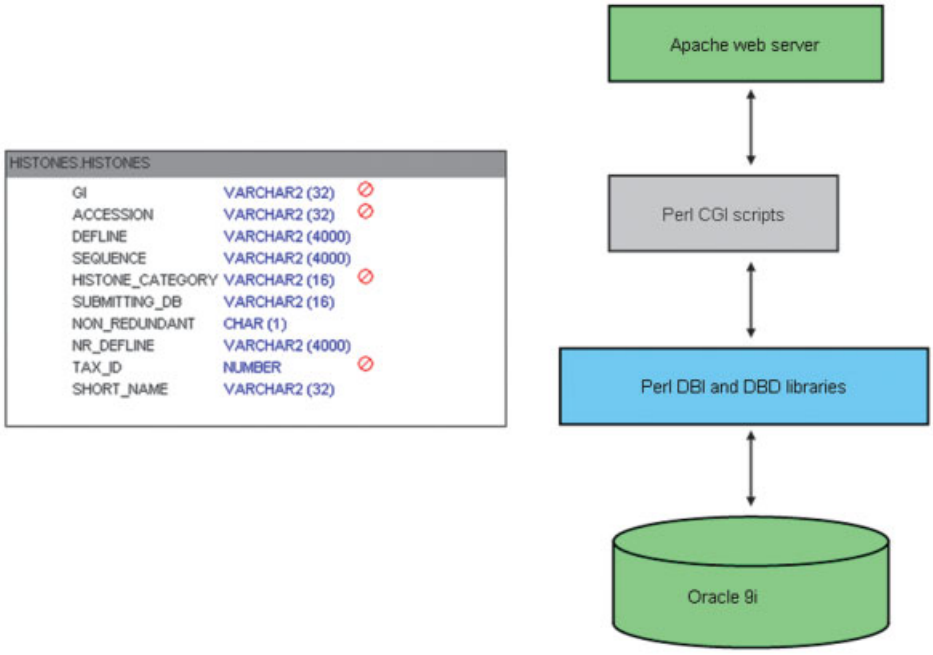


Fig. 1. Histone Database data model. The Histone Database stores selected information from the GenBank records, such as the GenBank unique identifier (GI), accession number, definition line, sequence string, and taxonomic identifier. The database front end is written in Perl, the data are stored in the relational database Oracle 9i, and are retrieved via Perl DBI and DBD libraries.

The obvious importance of histones to the overall structure of chromatin and in gene regulation led us to create and maintain a resource devoted to precisely and methodically cataloging these proteins. The Histone Database contains a collection of all histones and histone fold-containing proteins, with links to GenBank. The site also maintains a list of published three-dimensional structures for histones and histone fold-containing proteins, information on the human histone gene complement, multiple sequence alignments for each histone family, and information on posttranslational modifications.

## MATERIALS AND METHODS

### Database Schema and Implementation

The Histone Database consists of a series of Web interfaces written in the Perl programming language around a relational database. The use of object-oriented design methodologies and Perl modules that are both open source and developed in-house allows for flexibility and scalability. Several in-house modules, such as the ones that govern the results table display, are reused in other protein database applications. Web pages displaying data, such as the summary of contents, nonredundant sets, and search pages are dynamically generated using CGI.

The data are stored in a relational database schema using Oracle 9i (Fig. 1). Common data such as National Center for Biotechnology Information (NCBI) taxonomy identifiers are stored in different schemas and public synonyms are used to gather data across schemas.

Comments regarding the Web front-end are welcomed and encouraged.

### Data Sources and Generation

The protein databases searched were the NCBI's nonredundant database, which includes all nonredundant GenBank CDS translations, RefSeq proteins, Protein Data Bank (PDB), SwissProt, Protein Information Resource, and Protein Research Foundation. The collection of histones and histone fold-containing proteins was extended and revised, using PSI-BLAST to identify new proteins containing the motif.[13] We used each of the histones from the 2002 update as queries[14] for PSI-BLAST searches against the NCBI's nonredundant database. The PSI-BLAST searches were run to convergence with an *E*-value inclusion threshold of 0.01. For each histone family, multiple sequence alignments were generated using CLUSTALW[15] and MUSCLE.[16] The alignments are also available in PDF format and are color-coded to allow easy identification of amino acid variants. A summary table of the number of sequences found grouped by family and species represented in the database is provided (Table I).

Histone fold-containing proteins were identified as follows. We used the histone fold domain from each of the four core histone MUSCLE alignments (H2A, H2B, H3, and H4) as seeds for PSI-BLAST searches. The PSI-BLAST searches were run to convergence with an *E*-value inclusion threshold of 0.01. As a result, we were able to identify a total of 550 histone fold proteins and determine the specific contribution of each individual core histone profile toward the identification of these proteins (Table II).

**TABLE I. Histone Database Content**

| Core histone profile | Number of unique sequences | Increase since last update (%) | Number of species | Increase since last update (%) |
|---|---|---|---|---|
| H1/H5 | 248 | 39.3 | 93 | 31.0 |
| H2A | 323 | 93.4 | 123 | 70.8 |
| H2B | 289 | 62.4 | 114 | 60.6 |
| H3 | 397 | 257.7 | 857 | 138.7 |
| H4 | 121 | 68.1 | 144 | 45.5 |

**TABLE II. Nonhistone Proteins Containing Histone Folds Identified Using Combinations of Position Specific Scoring Matrices (PSSMs)**

| PSSM(s) used for identification | Sequences identified from profiles |
|---|---|
| H2A only | 279 |
| H2B only | 58 |
| H3 only | 48 |
| H4 only | 230 |
| H2A and H2B not H3 not H4 | 13 |
| H2B and H4 not H2A not H3 | 1 |
| H4 and H3 not H2A not H2B | 79 |
| H2A and H2B and H4 not H3 | 1 |
| H2A and H4 and H3 not H2B | 1 |
| H2A and H2B and H3 and H4 | 6 |

## RESULTS AND DISCUSSION

Since its inception in 1995, the Histone Database has been a valuable resource for researchers studying chromatin structure and function, as well as those actively involved in studying transcriptional regulation, where histone fold-containing proteins have a central role. Currently, the Histone Database contains entries that represent a total of 975 organisms. Sequences of the histone proteins and of nonhistones containing the histone fold are available in FASTA format. Additionally, a search engine is available for querying the database. The search engine has the ability to retrieve entries by protein family, organism, keyword, or based on a sequence pattern. The database also includes the three-dimensional structures for histone and histone fold-containing proteins in PDB; each structure has links to PDB and the Molecular Modelling Database, along with the protein name and source organism.

A number of histone fold-containing proteins have been identified among TATA-box binding protein-associated factors (TAFs) and transcription factors; however, the Ras activator Son of Sevenless remains the only cytoplasmic protein containing the histone fold motif. The structure of the histone fold domains from Son of Sevenless was recently determined.[12] The N-terminal structure of Son of Sevenless contains two histone folds that can be superimposed onto the H2A/H2B heterodimer with an root-mean-square deviation in $C_\alpha$ positions of only 1.2 Å. Interestingly, only the second histone fold was detected in a previous sequence analysis using PSI-BLAST searches.[4] However, position specific score matrices (PSSMs) can be constructed from structural alignments generated by VAST,[17] using other structural neighbors such as histone H2B and the transcription factor NF-Y. The structure-based alignment for the first domain of Son of Sevenless with histone H2B reveals a difference in the loop length between α-helices 2 and 3 (Fig. 2). When the gap in the alignment is included in the PSSM model, the first histone fold domain in Son of Sevenless is successfully identified. The function of the histone fold domains in Son of Sevenless is still unclear, but they are likely to be involved in the formation of higher-order oligomeric and/or heterotypic interactions with other histone fold-containing proteins.

Another newly identified histone fold-containing protein is the huntingtin-interacting protein M (HYPM; GenBank AAC26851); this protein is highly expressed in testis[18] and was originally found in a yeast two-hybrid screen using huntingtin as bait.[19] A multiple sequence alignment of HYPM with human, frog, and chicken histones H2A constructed using PSI-BLAST is shown in Figure 3. Interestingly, it has been shown that huntingtin interacts with Sp1 and TAFII130, causing changes in transcriptional regulation.[20] If huntingtin, HYPM, Sp1, or TAFII130 are part of the same complex, our findings suggest that HYPM could serve as a bridge between the complex and other unidentified histone fold-containing proteins.

As more and more sequence data continue to accumulate from the targeted sequencing of model genomes, it is interesting to speculate whether additional proteins that putatively contain the histone fold motif will be identified. Although it is difficult (if not impossible) to predict how many histone fold-containing proteins will be identified in the future, the constant refinement of methods such as those used in this study will lead to an improvement in our ability to identify these proteins with a high degree of confidence. In addition, an important computational challenge for the future will be not only to identify putative histone fold-containing proteins, but to use computational methods that will allow for the identification of these proteins' binding partners. Finally, we anticipate that future updates to this database will include a wider "evolutionary spread" of genomes as targeted sequencing efforts continue at an ever-increasing pace.

## CONCLUSIONS

The Histone Database is a comprehensive bioinformatic resource that compiles histone sequences and groups them into families. The Histone Database also maintains a collection of histone fold-containing sequences as well as three-dimensional structures available in PDB. The database is updated on a regular basis to continue to serve as a resource for structural and functional studies of histones and histone fold-containing proteins. Most importantly, information found in this database can be used to make novel biological discoveries, such as those regarding Son of Sevenless and the huntingtin-interacting protein M, described above.

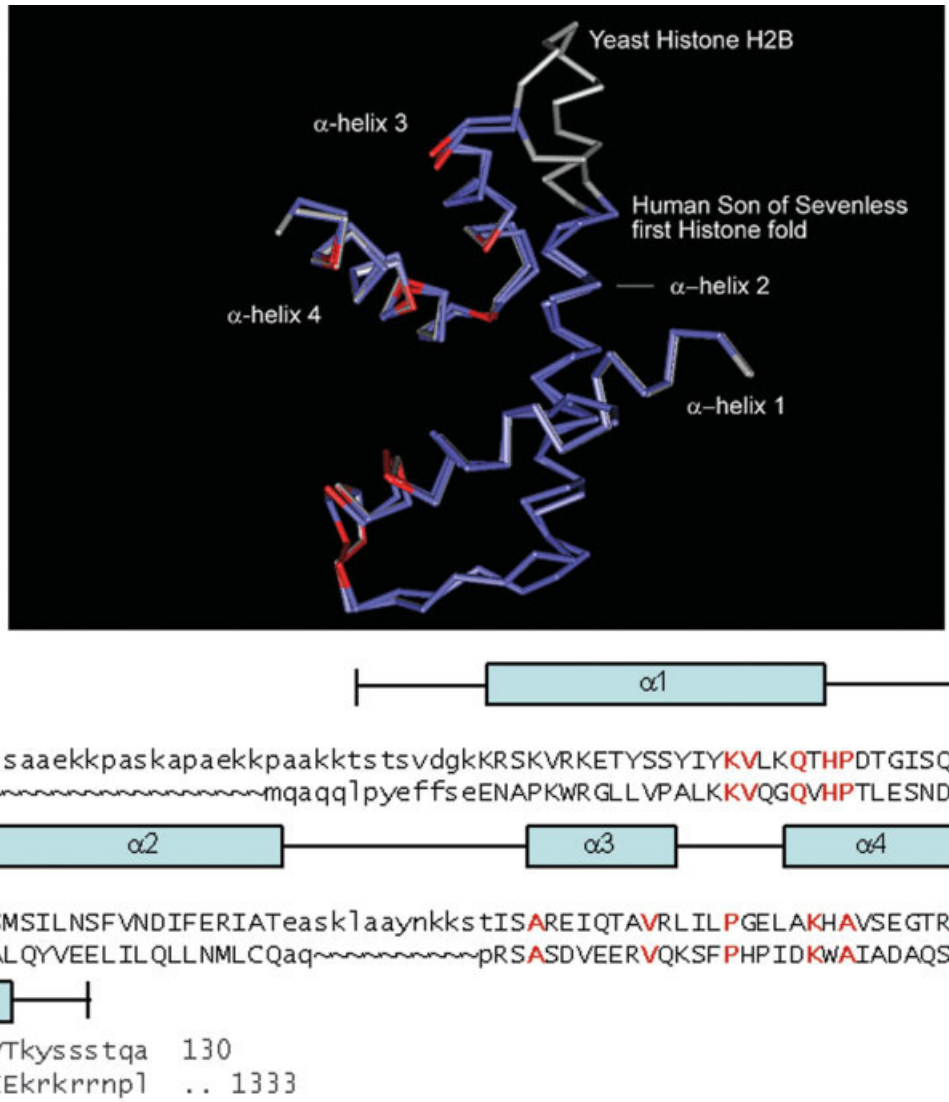


Fig. 2. Structure-based alignment of the first histone fold domain in human Son of Sevenless with yeast histone H2B. Yeast histone H2B (pdb|1ID3, chain D) aligned with the first histone fold domain present in the human ras activator Son of Sevenless (pdb|1Q9C, chain A). Secondary structural elements are represented above the sequence alignment and identical residues are colored in red.


Fig. 3. Multiple sequence alignment of three histone H2A members and the human huntingtin-interacting protein M (HYPM). Human, frog, and chicken histone H2A sequences aligned with the human HYPM protein. Secondary structural elements from the crystal structures of frog (pdb|1AOI, chain C) and chicken (pdb|2HIO, chain A) nucleosomes are represented above the sequence alignments and identical residues are colored in red; conserved residues in HYPM are uppercase.

## AVAILABILITY AND REQUIREMENTS

The Histone Database is freely available on the Web at http://research.nhgri.nih.gov/histones/. Studies that use the database should cite this article as the primary reference.

## ACKNOWLEDGMENTS

The authors are grateful to Julie D. Thompson for making the modifications to ClustalX that allowed us to generate PostScript files for larger paper sizes. We are also grateful to King Jordan for several helpful discussions. This study utilized the high-performance computational capabilities of the Biowulf PC/Linux cluster at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov/).

## REFERENCES

1. van Holde KE. Chromatin. New York: Springer-Verlag; 1988.
2. Eickbush TH, Moudrianakis EN. The histone core complex: an octamer assembled by two sets of protein-protein interactions. Biochemistry 1978;17(23):4955–4964.
3. Arents G, Moudrianakis EN. The histone fold: a ubiquitous architectural motif utilized in DNA compaction and protein dimerization. Proc Natl Acad Sci USA 1995;92(24):11170–11174.
4. Baxevanis AD, Arents G, Moudrianakis EN, Landsman D. A variety of DNA-binding and multimeric proteins contain the histone fold motif. Nucleic Acids Res 1995;23(14):2685–2691.
5. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature 1997;389(6648):251–260.
6. Bustin M, Catez F, Lim JH. The dynamics of histone H1 function in chromatin. Mol Cell 2005;17(5):617–620.
7. Khochbin S. Histone H1 diversity: bridging regulatory signals to linker histone function. Gene 2001;271(1):1–12.
8. Boeger H, Griesenbeck J, Strattan JS, Kornberg RD. Nucleosomes unfold completely at a transcriptionally active promoter. Mol Cell 2003;11(6):1587–1598.
9. Boeger H, Griesenbeck J, Strattan JS, Kornberg RD. Removal of promoter nucleosomes by disassembly rather than sliding in vivo. Mol Cell 2004;14(5):667–673.
10. Jenuwein T, Allis CD. Translating the histone code. Science 2001;293(5532):1074–1080.
11. Malik HS, Henikoff S. Phylogenomics of the nucleosome. Nat Struct Biol 2003;10(11):882–891.
12. Sondermann H, Soisson SM, Bar-Sagi D, Kuriyan J. Tandem histone folds in the structure of the N-terminal segment of the ras activator Son of Sevenless. Structure (Camb) 2003;11(12):1583–1593.
13. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389–3402.
14. Sullivan S, Sink DW, Trout KL, et al. The Histone Database. Nucleic Acids Res 2002;30(1):341–342.
15. Chenna R, Sugawara H, Koike T, et al. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 2003;31(13):3497–3500.
16. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32(5):1792–1797.
17. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6(3):377–385.
18. Su AI, Wiltshire T, Batalov S, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA 2004;101(16):6062–6067.
19. Faber PW, Barnes GT, Srinidhi J, Chen J, Gusella JF, MacDonald ME. Huntingtin interacts with a family of WW domain proteins. Hum Mol Genet 1998;7(9):1463–1474.
20. Dunah AW, Jeong H, Griffin A, et al. Sp1 and TAFII130 transcriptional activity disrupted in early Huntington's disease. Science 2002;296(5576):2238–2243.